Evaluation of ChatGPT for Myocardial Infarction Health Information: Expert Accuracy Assessment and User Trust Survey

Eman Elsheikh[1,2], Abdullah Ahmed Aladsani[3], Yousef Abdullatif Aljohar[3], Ghala Abdulaziz Alzahrani[3], Ali Mohammed Nasser Alsadah[3], Maryam Abdulmohsen Almuhaysh[3], Walah Sami Rashid alowaynan[3], Rehab H. Werida[4]

[1] Department of Cardiology, College of Medicine, Tanta University, Tanta, Egypt
[2] Internal Medicine Department, College of Medicine, King Faisal University, Alahsa, Saudi Arabia
[3] College of Medicine, King Faisal University, Alahsa, Saudi Arabia
[4] Clinical Pharmacy & Pharmacy Practice Department, Faculty of Pharmacy, Damanhur University, Egypt.

Abstract

Background: Artificial intelligence chatbots, particularly ChatGPT, have emerged as increasingly popular sources of health information for the general public. However, concerns persist regarding the accuracy, safety, and appropriateness of AI-generated medical advice, especially for life-threatening conditions such as myocardial infarction.

Objectives: This study aimed to evaluate the accuracy, completeness, and safety of ChatGPT-generated responses and information to public questions about heart attacks, assess user perceptions and trust, and compare AI-generated content with expert-validated sources.

Methods: A cross-sectional study was conducted between January and March 2026. A total of 73 commonly asked heart attack-related questions were submitted to ChatGPT (GPT-4), and responses were independently evaluated by two expert reviewers and one external reviewer using standardized rubric assessing accuracy, completeness, clarity, safety, and tone. Inter-rater reliability was assessed using intraclass correlation coefficients. In parallel, a survey of 352 participants evaluated user perceptions, trust, and behavioral use of ChatGPT for medical information. Statistical analyses included descriptive statistics, group comparisons, correlation analyses, and multivariable regression models to identify predictors of trust and satisfaction.

Results: Expert reviewers assigned high ratings for accuracy (mean: 4.62 ± 0.62 and 4.48 ± 0.63) and safety (mean: 4.78 ± 0.51 and 4.05 ± 0.50), with substantial inter-rater agreement (ICC = 0.788). In contrast, the external reviewer documented considerably lower completeness ratings (2.59 ± 0.57), revealing deficiencies in information coverage. Remarkably, 40.6% of respondents indicated they had followed ChatGPT's medical recommendations without seeking professional consultation, while 17.6% reported depending on the chatbot during urgent medical situations. Levels of user trust and satisfaction were moderate, showing significant correlations with perceived accuracy, usage frequency, and healthcare-related educational background.

Conclusions: Expert assessments confirmed that ChatGPT generally provides accurate and safe cardiovascular health information; nevertheless, considerable inter-reviewer variability—especially the markedly lower completeness scores from the external evaluator—reveals significant gaps in content depth. These results indicate that ChatGPT functions more appropriately as an adjunctive educational resource rather than a replacement for professional medical consultation, emphasizing the necessity for enhanced safety mechanisms, clearer user instructions, and standardized content protocols when addressing high-stakes health topics. ChatGPT should be considered an adjunctive resource rather than a substitute for urgent professional medical assessment in suspected heart attack cases.

Keywords: ChatGPT, artificial intelligence, heart attack, myocardial infarction, health literacy, patient education, digital healthIntroduction

How to cite this article: Elsheikh E, Aladsani AA, Aljohar YA, Alzahrani GA, Alsadah AMN, Almuhaysh MA, alowaynan WSR, Werida RH. Evaluation of ChatGPT for Myocardial Infarction Health Information: Expert Accuracy Assessment and User Trust Survey. Int J Drug Deliv Technol. 2026;16(70s): 384-411. DOI: 10.25258/ijddt.16.70s.39

The rapid proliferation of artificial intelligence technologies has fundamentally transformed how individuals' access and consume health information. Large language models such as ChatGPT have emerged as widely utilized tools capable of generating human-like responses across diverse domains, including medicine and healthcare. Since their introduction, these systems have gained substantial attention among both healthcare professionals and the general public, with many individuals

increasingly relying on them for immediate medical guidance. This growing dependence reflects a broader shift toward digital, on-demand health information seeking, raising important concerns about the reliability, accuracy, and safety of AI-generated medical content (1,2).

Cardiovascular diseases remain the leading cause of mortality worldwide, with conditions such as myocardial infarction representing critical medical emergencies that require rapid recognition and timely intervention. Public awareness of symptoms, risk factors, and appropriate responses plays a vital role in improving survival outcomes. However, gaps in knowledge and delays in seeking care continue to persist, potentially exacerbated by misinformation or incomplete understanding. As AI tools become increasingly integrated into health information-seeking behaviours, their role in shaping public understanding of life-threatening conditions such as heart attacks warrants careful evaluation (3,4).

Traditionally, health information has been disseminated through healthcare providers, printed educational materials, and reputable online platforms. The emergence of AI chatbots introduces a new paradigm characterized by rapid accessibility, conversational interaction, and perceived personalization. While these features enhance user engagement, they also raise concerns regarding the consistency, completeness, and clinical appropriateness of the information provided. Studies have shown that individuals may rely on AI-generated responses before consulting healthcare professionals, highlighting the potential impact of such tools on decision-making, particularly in urgent or high-risk scenarios (5,6).

The performance of AI systems in medical contexts has demonstrated variability across different evaluation domains. While some research indicates high levels of accuracy and appropriateness in structured or guideline-based questions, other studies have identified limitations, including incomplete responses, variability in output, and challenges in addressing complex or emergency-related situations. These inconsistencies emphasize the need to evaluate not only the accuracy of AI-generated information but also its completeness and safety, particularly when applied to critical conditions where inappropriate advice may have serious consequences (7,8).

Additionally, factors such as readability, user trust, and health literacy significantly influence the effectiveness of health communication. Although AI tools have demonstrated potential in simplifying complex medical information, variations in clarity and actionability may affect user comprehension and subsequent behaviour. Given that inadequate health literacy is associated with poorer health outcomes and increased healthcare utilization, ensuring that AI-generated responses are both understandable and reliable is essential. Furthermore, concerns regarding patient safety, including the risk of delayed diagnosis or inappropriate self-management, underscore the importance of critically assessing these technologies within real-world contexts (9,10).

To bridge these knowledge gaps, this investigation was structured as a mixed-method evaluation that integrates systematic assessment of AI-generated medical content with analysis of user attitudes and behaviours. The primary outcome focused on determining the accuracy of ChatGPT responses to heart attack-related inquiries through expert evaluation. Secondary outcomes encompassed completeness, clarity (readability), safety, and communication tone of the generated responses, alongside user trust levels, satisfaction ratings, and patterns of behavioural dependence on ChatGPT for health-related decisions.

Methods

Study Design

This investigation utilized a cross-sectional approach comprising two interconnected elements: (1) systematic validation of ChatGPT-generated content through expert assessment, and (2) evaluation of user attitudes, confidence levels, and utilization behaviour. The methodology was designed to differentiate between objective quality indicators (accuracy, completeness, clarity, safety, and tone) and subjective participant-reported measures (trust, satisfaction, and behavioural dependence).

Participant Recruitment and Sample

Survey participants were enrolled using convenience sampling through social media channels and professional networks. Eligibility requirements included being 18 years of age or older and possessing the ability to comprehend and complete the survey. No exclusion criteria were implemented to ensure maximum sample diversity. The final sample consisted of 352

respondents. Notably, a substantial proportion of participants (75.3%) possessed healthcare-related training, attributable to recruitment through professional networks alongside public platforms. Consequently, the sample reflects a population with above-average health literacy rather than exclusively representing the general public. To address this characteristic, medical background was incorporated as a primary stratification factor in all subsequent statistical analyses.

Survey Instrument

The survey instrument was developed based on a review of existing literature on AI use in healthcare and adapted to the specific context of ChatGPT utilization for medical information. The questionnaire comprised several sections including demographic information (age, gender, education level, medical background), prior experience with ChatGPT and AI chatbots, frequency and nature of health-related queries, and attitudes toward AI-generated medical information.

Attitude items were measured using 5-point Likert scales ranging from 1 (strongly disagree) to 5 (strongly agree). These items assessed beliefs regarding ChatGPT's accuracy, ease of understanding, balance, convenience compared to physician consultation, trustworthiness relative to traditional search engines, confidence in AI responses, currency of information, and perception of conflicting answers. Additional items assessed verification behaviours, awareness of potential errors, opinions on warning inclusion, overall satisfaction, and recommendation intentions.

ChatGPT Response Evaluation

A curated set of 73 questions representing commonly asked public queries about heart attacks was developed from frequently asked questions sections of authoritative health websites, including the American Heart Association, British Heart Foundation, and National Health Service. Questions covered domains including symptoms recognition, risk factors, emergency response, treatment options, recovery, and prevention.

Questions were organized into predetermined thematic categories—symptom identification, risk factors, emergency protocols, therapeutic approaches, recovery processes, and preventive strategies—to guarantee thorough representation of essential myocardial infarction educational content. Two clinical specialists (in cardiology and internal medicine) independently assessed the question set for appropriateness, linguistic precision, and alignment with typical patient concerns, thereby strengthening content validity. Consensus-based revisions were implemented to enhance question clarity and correspondence with authentic patient queries. Prior to data analysis, inter-rater concordance regarding question suitability and categorical assignment was evaluated, with any disagreements reconciled through collaborative discussion.

ChatGPT Query Protocol and Reproducibility Settings

All questions were submitted to ChatGPT using the GPT-4 model accessed via web interface in February 2026. To maximize reproducibility, queries were conducted under uniform conditions with standardized phrasing and consistent prompt formatting. Each question was submitted as an independent, isolated prompt without contextual modifications, follow-up queries, clarification requests, or iterative revisions. A single-run methodology was employed, with no response regeneration or selection from multiple outputs. Default system configurations were maintained without manual adjustment of parameters such as temperature or response variability. Responses were documented verbatim immediately upon generation to prevent temporal inconsistencies. No system-level prompt modifications were implemented beyond standard interface settings, ensuring outputs accurately represent typical user interactions.

Expert Evaluation Process

Two expert reviewers with backgrounds in cardiology and clinical pharmacology independently evaluated each ChatGPT response using a standardized rubric. The evaluation criteria included accuracy (medical correctness compared to established guidelines), completeness (coverage of necessary information points), clarity (understandability for lay audiences), safety (absence of harmful or misleading advice), and tone (supportive and appropriate communication style). Each criterion was rated on a 5-point scale, with higher scores indicating better performance. A composite quality score was calculated by summing the five individual domain scores. To enhance the validity of accuracy assessment, ChatGPT responses were evaluated against established clinical reference

standards derived from authoritative guidelines, including the American Heart Association (AHA) and European Society of Cardiology (ESC) recommendations on myocardial infarction. Reviewers were instructed to consider alignment with guideline-based definitions, symptom recognition, emergency response recommendations, and management principles when assigning accuracy and completeness scores.

An external independent reviewer also evaluated all responses using the same criteria to provide an additional perspective and assess consistency of findings. Furthermore, a sample of general public raters evaluated the responses on dimensions of perceived trustworthiness, helpfulness, and ease of understanding.

Comparison ratings were also obtained by having evaluators assess relevant content from established comparator sites (American Heart Association and British Heart Foundation websites) on accuracy and completeness dimensions.

Ethical Considerations

The study protocol was approved by King Faisal University's Medical Research Centre (KFU-REC-2025-NOV–ETHICS313) prior to data collection. Informed consent was obtained from all participants before questionnaire completion, with surveys administered anonymously. No patient or clinical data were utilized in this investigation. ChatGPT outputs were de-identified and stored securely in compliance with data protection regulations. All collected information remains confidential and is used exclusively for research purposes.

Statistical Analysis

Statistical analyses were performed using SPSS version 28 (IBM Corporation, Armonk, NY, USA). The Shapiro-Wilk test and visual inspection of histograms were used to evaluate the normality of data distributions. Continuous variables following normal distributions were presented as means and standard deviations and analysed using analysis of variance with Tukey post hoc tests for multiple comparisons. Non-normally distributed continuous variables were presented as medians and interquartile ranges and analysed using the Kruskal-Wallis test with Mann-Whitney U tests for pairwise comparisons. Categorical variables were presented as frequencies and percentages and analysed using chi-square tests.

Spearman correlation coefficients were calculated to assess relationships between continuous variables. Multiple regression analyses were conducted to identify predictors of key outcomes including general public trust, satisfaction with ChatGPT as a medical information source, and intention to recommend ChatGPT. Inter-rater reliability was assessed using intraclass correlation coefficients with two-way random effects models for absolute agreement. ICC values were interpreted according to established thresholds: values below 0.50 signified poor agreement, values between 0.50 and 0.75 indicated moderate agreement, and values exceeding 0.75 reflected good agreement. Reviewer concordance was additionally examined through Bland-Altman analysis. Statistical significance was defined as a two-tailed p-value below 0.05.

Analytical procedures were organized according to predetermined primary and secondary research objectives. The primary analysis cantered on accuracy assessment of ChatGPT responses based on expert ratings. Secondary analyses encompassed evaluation of completeness, clarity, safety, and tone, in addition to inter-reviewer comparisons. Supplementary exploratory analyses investigated relationships between participant demographics and ChatGPT perceptions, including trust, satisfaction, and usage behaviours. Multivariable regression modelling identified independent predictors of principal user-cantered outcomes.

Results:

❖ **Participant Characteristics**

The demographic characteristics of the studied participants are presented in **Table 1**. The current study included 352 participants, of whom 88 (25%) were males and 264 (75%) were females. The mean age was 38.55 ± 13.02 years. Regarding educational level, the majority of participants held a bachelor's degree (56.8%), followed by high school education (22.2%) and postgraduate level (12.2%). Smaller proportions had diploma (3.7%), PhD (2.3%), assistant professor (0.9%), and master's degree (0.3%). Most participants (75.3%) had a medical background.

Table 1: Demographic data of the studied participants

	Total (n=352)

Age (years)		38.55 ± 13.02
Gender	**Male**	88 (25%)
	Female	264 (75%)
Education level	**Intermediate education**	6 (1.7%)
	High school	78 (22.2%)
	Bachelor	200 (56.8%)
	Postgraduate	43 (12.2%)
	Diploma	13 (3.7%)
	Master	1 (0.3%)
	PhD	8 (2.3%)
	Assistant professor	3 (0.9%)
Medical background		265 (75.3%)

❖ **User Awareness, Usage, and Perceptions of ChatGPT**

The level of awareness, use, and perceptions of ChatGPT is shown in **Table 2**. A total of 320 participants (90.9%) had heard of ChatGPT or AI chatbots, and 254 (72.2%) reported using ChatGPT to search for medical or health information. Regarding frequency of use, 15.6% used it daily, 38.6% occasionally, 15.3% weekly, and 30.4% rarely. The most commonly searched topics included symptoms, medications, and treatment options, either individually or in combination. Social media (44.3%) and friends (35.2%) were the main sources of initial exposure to ChatGPT.

Participants' perceptions showed moderate agreement regarding the accuracy of ChatGPT (3.03 ± 1.2), while higher agreement was reported for ease of understanding (3.88 ± 1.23). Scores for perceived balance (3.42 ± 1.26), trust compared to Google (3.2 ± 1.42), confidence in language (3.31 ± 1.28), and use of up-to-date information (3.31 ± 1.19) were moderate. A total of 143 participants (40.6%) reported acting on ChatGPT's medical advice without consulting a doctor, and 62 (17.6%) reported relying on it in emergency situations. Participants demonstrated relatively high awareness of potential factual errors (4.07 ± 1.16) and strong agreement on the need for clear warnings in AI-generated medical information (4.24 ± 1.13).

Table 2: AI knowledge level of the studied participants

		Total (n=352)
Ever heard of ChatGPT or AI chatbots		320 (90.9%)
Ever used ChatGPT to search for medical or health information		254 (72.2%)
How often they use ChatGPT for health-related questions	**Daily**	55 (15.6%)
	Occasionally	136 (38.6%)
	Rarely	107 (30.4%)
	Weekly	54 (15.3%)
Kind of topics they usually ask about	**Symptoms**	33 (9.4%)
	Medication	19 (5.4%)
	Treatment Options	11 (3.1%)
	Nutrition	4 (1.1%)
	Exercise	11 (3.1%)
	Study and research related	11 (3.1%)
	Others	20 (5.7%)
	Not used	3 (0.9%)
	Symptoms & medications	25 (7.1%)
	Symptoms & treatment options	24 (6.8%)
	Symptoms & nutrition	3 (0.9%)
	Symptoms & exercise	3 (0.9%)
	Symptoms & others	3 (0.9%)
	Symptoms, medications, treatment options & nutrition	10 (2.8%)
	Symptoms, medications, treatment options & exercise	1 (0.3%)
	Symptoms, medications & nutrition	8 (2.3%)
	Symptoms, medications, nutrition & exercise	7 (2%)

	Symptoms, medications, nutrition, exercise & others	2 (0.6%)
	Symptoms, medications & exercise	8 (2.3%)
	Symptoms, medications & others	4 (1.1%)
	Symptoms, medications & study and research related	2 (0.6%)
	Symptoms, treatment options & nutrition	7 (2%)
	Symptoms, treatment options, nutrition & exercise	1 (0.3%)
	Symptoms, treatment options & exercise	6 (1.7%)
	Symptoms, treatment options, exercise & others	3 (0.9%)
	Symptoms, nutrition & exercise	1 (0.3%)
	Symptoms, nutrition & others	3 (0.9%)
	Symptoms, medications, nutrition, exercise & others	1 (0.3%)
	Symptoms, medications & treatment options	35 (9.9%)
	Symptoms, medications, treatment options, nutrition & exercise	41 (11.6%)
	Symptoms, medications, treatment options, nutrition, exercise & others	3 (0.9%)
	Symptoms, medications, treatment options, nutrition, exercise & study and research related	3 (0.9%)
	Medications & treatment options	11 (3.1%)
	Medications & nutrition	3 (0.9%)
	Medications, treatment options & nutrition	4 (1.1%)
	Medications, treatment options, nutrition & exercise	2 (0.6%)
	Medications, nutrition & exercise	2 (0.6%)
	Nutrition & exercise	7 (2%)
	Treatment options, nutrition & exercise	3 (0.9%)
	Treatment options & exercise	4 (1.1%)
The way of the first learn about ChatGPT	**No**	6 (1.7%)
	Children	9 (2.6%)
	Family	7 (2%)
	Friends	124 (35.2%)
	News	20 (5.7%)
	Social media	156 (44.3%)
	Social media and friends	1 (0.3%)
	Trends	3 (0.9%)
	University	20 (5.7%)
	All of the above	6 (1.7%)
Believing that ChatGPT provides accurate medical information		3.03 ± 1.2
		3 (2-4)

ChatGPT's explanations are easy to understand		3.88 ± 1.23
		4 (3-5)
ChatGPT gives balanced and unbiased health advice		3.42 ± 1.26
		3 (3-5)
ChatGPT is more convenient than visiting a doctor for minor issues		2.88 ± 1.5
		3 (1-4)
Trusting ChatGPT's medical advice as much as information from Google		3.2 ± 1.42
		3 (2-5)
ChatGPT's language makes me feel confident in its answers		3.31 ± 1.28
		3 (3-4)
Believing that ChatGPT always uses up-to-date medical information		3.31 ± 1.19
		3 (3-4)
Feeling ChatGPT sometimes gives conflicting or confusing answers		3.31 ± 1.21
		3 (2-4)
Ever acted on ChatGPT's medical advice without consulting a doctor		143 (40.6%)
Relying on ChatGPT in an emergency	**Yes**	62 (17.6%)
	No	146 (41.5%)
	Not sure	144 (40.9%)
Always verifying ChatGPT's answers with official health websites or professionals		3.62 ± 1.25
		4 (3-5)
Aware that ChatGPT can sometimes make factual errors		4.07 ± 1.16
		5 (3-5)
Thinking ChatGPT should include clear warnings when giving medical information		4.24 ± 1.13
		5 (4-5)
Satisfaction with ChatGPT as a medical-information source	**Very unsatisfied**	25 (7.1%)
	Unsatisfied	73 (20.7%)
	Neutral	119 (33.8%)
	Satisfied	107 (30.4%)
	Very Satisfied	28 (8%)
Recommending ChatGPT to friends or family for health advice	**Yes**	220 (62.5%)
	No	78 (22.2%)
	Not sure	54 (15.3%)

Figure 1 illustrates the distribution of participant satisfaction with ChatGPT as a medical information source. The largest segment of respondents expressed neutral to moderate satisfaction levels. Although a considerable proportion reported positive experiences, a meaningful percentage expressed dissatisfaction, demonstrating heterogeneity in user experiences and the absence of consistently favourable perceptions.

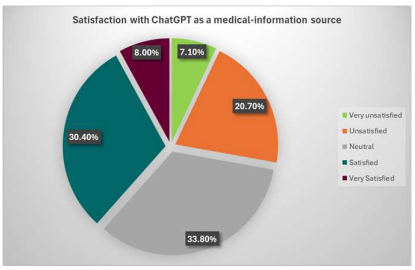

Figure 1: Satisfaction with ChatGPT as a medical-information source

Figure 2 reveals that most participants expressed willingness to recommend ChatGPT for health-related inquiries. Nevertheless, the substantial proportion of respondents who remained uncertain or declined to recommend the platform indicates persistent concerns about its dependability and credibility.

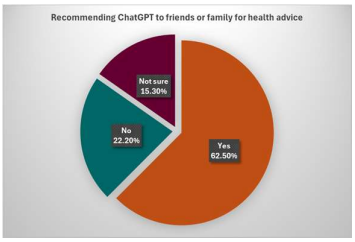

Figure 2: Recommendation of ChatGPT to friends or family for health advice

Primary Outcome: Accuracy of ChatGPT Responses

Descriptive statistics of expert evaluation scores for accuracy are presented in Table 3. Expert reviewer 1 demonstrated high mean accuracy scores (4.62 ± 0.62), while Expert reviewer 2 showed slightly lower but still high scores (4.48 ± 0.63). The external reviewer reported comparatively lower accuracy scores (3.89 ± 0.49). Comparator websites demonstrated moderate performance in accuracy (4.00 ± 0.24).

Table 3: Accuracy Scores from Expert Evaluation of ChatGPT Responses (N=73 questions)

Evaluator	Mean ± SD	Median (IQR)
Expert Reviewer 1	4.62 ± 0.62	5 (4–5)
Expert Reviewer 2	4.48 ± 0.63	5 (4–5)
External Independent Reviewer	3.89 ± 0.49	4 (4–4)
Comparator Sites (AHA/BHF)	4.00 ± 0.24	4 (4–4)

Scores rated on 5-point scale (1=lowest, 5=highest); SD=standard deviation; IQR=interquartile range; AHA=American Heart Association; BHF=British Heart Foundation

Reliability analysis is presented in **Table 4**, demonstrating good agreement between expert reviewers 1 and 2 (ICC = 0.788), fair agreement between expert 1 and the external reviewer (ICC = 0.544), and good agreement between expert 2 and the external reviewer (ICC = 0.850), all statistically significant (P < 0.001).

Table 4: Intraclass Correlation Coefficients for Inter-Rater Reliability

Reviewer Comparison	ICC	95% CI	F Value	df1	df2	P-value
Expert 1 vs. Expert 2	0.788	0.646–0.871	6.670	72	288	<0.001*
Expert 1 vs. External Reviewer	0.544	0.213–0.734	4.459	72	288	<0.001*
Expert 2 vs. External Reviewer	0.850	0.788–0.898	6.670	72	288	<0.001*

*ICC=Intraclass Correlation Coefficient; CI=Confidence Interval; df=degrees of freedom; *Statistically significant (P<0.05)*

Secondary Outcomes: Content Quality Assessment

Descriptive statistics of expert evaluation scores for secondary outcomes are presented in Table 5. Expert reviewer 1 demonstrated high mean scores across domains, including completeness (3.92 ± 0.52), clarity (4.01 ± 0.26), safety (4.78 ± 0.51), and tone (4.99 ± 0.12), with an overall quality score of 22.32 ± 1.61. Expert reviewer 2 showed slightly lower but still high scores, while the external reviewer reported comparatively lower scores across all domains. Comparator websites demonstrated moderate performance in completeness (3.97 ± 0.50). Public evaluation indicated moderate trust (3.26 ± 0.53), high helpfulness (3.92 ± 0.36), and high ease of understanding (4.33 ± 0.73).

Table 5: Comprehensive Expert Evaluation Scores for ChatGPT Responses (N=73 questions)

Evaluator	Domain	Mean ± SD	Median (IQR)
Expert Reviewer 1	Accuracy	4.62 ± 0.62	5 (4–5)
	Completeness	3.92 ± 0.52	4 (4–4)
	Clarity	4.01 ± 0.26	4 (4–4)
	Safety	4.78 ± 0.51	5 (5–5)
	Tone	4.99 ± 0.12	5 (5–5)
	Total Quality Score	**22.32 ± 1.61**	**23 (22–23)**
Expert Reviewer 2	Accuracy	4.48 ± 0.63	5 (4–5)
	Completeness	3.95 ± 0.60	4 (4–4)
	Clarity	3.73 ± 0.53	4 (3–4)
	Safety	4.05 ± 0.50	4 (4–4)
	Tone	4.08 ± 0.46	4 (4–4)
	Total Quality Score	**20.29 ± 2.16**	**20 (19–21)**

External Reviewer	Accuracy	3.89 ± 0.49	4 (4–4)
	Completeness	2.59 ± 0.57	3 (2–3)
	Clarity	3.40 ± 0.78	4 (3–4)
	Safety	3.86 ± 0.61	4 (3–4)
	Tone	3.33 ± 0.53	3 (3–4)
	Total Quality Score	**17.07 ± 1.16**	**17 (17–17)**
Comparator Sites	Accuracy	4.00 ± 0.24	4 (4–4)
	Completeness	3.97 ± 0.50	4 (4–4)
General Public	Trustworthiness	3.26 ± 0.53	3 (3–3)
	Helpfulness	3.92 ± 0.36	4 (4–4)
	Ease of Understanding	4.33 ± 0.73	4 (4–5)

- *Scores rated on 5-point scale (1=lowest, 5=highest); Total Quality Score range: 5–25; SD=standard deviation; IQR=interquartile range*

Reviewer Comparison Analysis

Comparative analysis between reviewers is shown in **Table 6**, where significant differences were observed across all domains and total quality scores ($P < 0.05$). Accuracy and completeness scores were significantly higher among expert reviewers 1 and 2 compared to the external reviewer ($P < 0.05$), with no significant difference between the two experts. Clarity, tone, and overall quality scores were highest for expert reviewer 1, followed by expert reviewer 2, and lowest for the external reviewer. Safety scores were significantly higher for expert reviewer 1 compared to both other reviewers ($P < 0.001$).

Table 6: Comparison of expert evaluation scores among reviewers

Domain	Expert 1 (n=73)	Expert 2 (n=73)	External Reviewer (n=73)	P-value	Post-hoc Comparisons
Accuracy	4.62 ± 0.62	4.48 ± 0.63	3.89 ± 0.49	<0.001*	P1=0.329; P2<0.001*; P3<0.001*
	5 (4–5)	5 (4–5)	4 (4–4)		
Completeness	3.92 ± 0.52	3.95 ± 0.60	2.59 ± 0.57	<0.001*	P1=0.954; P2<0.001*; P3<0.001*
	4 (4–4)	4 (4–4)	3 (2–3)		
Clarity	4.01 ± 0.26	3.73 ± 0.53	3.40 ± 0.78	<0.001*	P1=0.007*; P2<0.001*; P3=0.002*
	4 (4–4)	4 (3–4)	4 (3–4)		
Safety	4.78 ± 0.51	4.05 ± 0.50	3.86 ± 0.61	<0.001*	P1<0.001*; P2<0.001*; P3=0.083
	5 (5–5)	4 (4–4)	4 (3–4)		
Tone	4.99 ± 0.12	4.08 ± 0.46	3.33 ± 0.53	<0.001*	P1<0.001*; P2<0.001*; P3<0.001*
	5 (5–5)	4 (4–4)	3 (3–4)		
Total Quality Score	22.32 ± 1.61	20.29 ± 2.16	17.07 ± 1.16	<0.001*	P1<0.001*; P2<0.001*; P3<0.001*

	23 (22 – 23)	20 (19 – 21)	17 (17–17)		

*Data presented as mean ± SD and median (IQR); P1=Expert 1 vs. Expert 2; P2=Expert 1 vs. External; P3=Expert 2 vs. External; *Statistically significant (P<0.05)*

The Bland–Altman plot (Figure 3) further demonstrates the concordance between expert reviewers 1 and 2. The analysis reveals satisfactory agreement, with the majority of score differences falling within acceptable limits and no evident systematic bias across the scoring range, confirming the reliability of expert assessments.

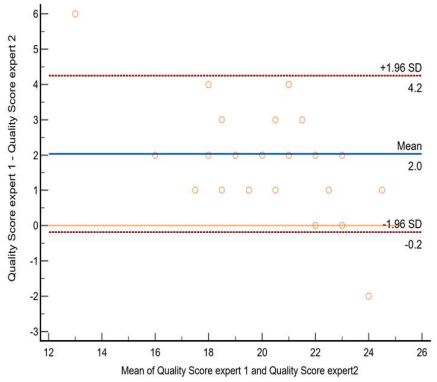

Figure 3: Bland–Altman plot for agreement between expert reviewers 1 and 2

Correlation analysis (**Table 7**) demonstrated a strong positive correlation between the quality scores of expert reviewers 1 and 2 (r = 0.860, P < 0.001), while no significant correlations were observed between either expert and the external reviewer.

Table 7: Correlation between quality scores of expert reviewers

		Quality score expert 1	Quality score expert 2	External expert 3 total score
Quality score expert 1	**r**	--	0.860	0.063
	P	--	**<0.001***	0.599
Quality score expert 2	**r**	0.860	--	0.086
	P	**<0.001***	--	0.468
Quality score of external expert 3	**r**	0.063	0.086	--
	P	0.599	0.468	--

Factors Associated with ChatGPT Use

As shown in **Table 8**, there was a statistically significant association between the use of ChatGPT for medical or health information and age (P < 0.001), with younger participants more likely to use it. Significant associations were also observed with having a medical background and prior awareness of ChatGPT (P < 0.001 for both). Additionally, significant relationships were found between ChatGPT use and frequency of use, types of topics searched, and initial source of awareness (P < 0.001 for all).

Participants who used ChatGPT demonstrated significantly higher scores in several perception domains, including perceived accuracy, clarity, balance, trust, confidence, and belief in up-to-date information (P < 0.05). They were also significantly more likely to report acting on ChatGPT's medical advice without consulting a doctor (P < 0.001). Furthermore, satisfaction with ChatGPT and likelihood of recommending it were significantly higher among users (P < 0.001). No significant associations were found between ChatGPT use and gender, educational level, perceived convenience over visiting a doctor, verification behaviour, awareness of errors, or the need for warnings.

Table 8: Relation between the use of ChatGPT and demographic data and AI knowledge level

		Use of ChatGPT		P value
		Yes (n=254)	No (n=98)	
Age (years)		35.49 ± 12.6	46.5 ± 10.55	**<0.001***
Gender	**Male**	69 (27.2%)	19 (19.4%)	0.131

	Female	185 (72.8%)	79 (80.6%)	
Education level	**Intermediate education**	3 (1.2%)	3 (3.1%)	0.432
	High school	53 (20.9%)	25 (25.5%)	
	Bachelor	149 (58.7%)	51 (52%)	
	Postgraduate	31 (12.2%)	12 (12.2%)	
	Diploma	10 (3.9%)	3 (3.1%)	
	Master	0 (0%)	1 (1%)	
	PhD	5 (2%)	3 (3.1%)	
	Assistant professor	3 (1.2%)	0 (0%)	
Medical background		210 (82.7%)	55 (56.1%)	**<0.001***
Ever heard of ChatGPT or AI chatbots		253 (99.6%)	67 (68.4%)	**<0.001***
How often they use ChatGPT for health-related questions	**Daily**	48 (18.9%)	7 (7.1%)	**<0.001***
	Occasionally	106 (41.7%)	30 (30.6%)	
	Rarely	55 (21.7%)	52 (53.1%)	
	Weekly	45 (17.7%)	9 (9.2%)	
Kind of topics they usually ask about	**Symptoms**	25 (9.8%)	8 (8.2%)	**<0.001***
	Medication	13 (5.1%)	6 (6.1%)	
	Treatment Options	7 (2.8%)	4 (4.1%)	
	Nutrition	2 (0.8%)	2 (2%)	
	Exercise	1 (0.4%)	10 (10.2%)	
	Study and research related	5 (2%)	6 (6.1%)	
	Others	4 (1.6%)	16 (16.3%)	
	Not used	0 (0%)	3 (3.1%)	
	Symptoms & medications	23 (9.1%)	2 (2%)	
	Symptoms & treatment options	24 (9.4%)	0 (0%)	
	Symptoms & nutrition	0 (0%)	3 (3.1%)	
	Symptoms & exercise	2 (0.8%)	1 (1%)	
	Symptoms & others	3 (1.2%)	0 (0%)	
	Symptoms, medications, treatment options & nutrition	9 (3.5%)	1 (1%)	
	Symptoms, medications, treatment options & exercise	1 (0.4%)	0 (0%)	
	Symptoms, medications & nutrition	6 (2.4%)	2 (2%)	

	Symptoms, medications, nutrition & exercise	6 (2.4 %)	1 (1%)	
	Symptoms, medications, nutrition, exercise & others	2 (0.8 %)	0 (0%)	
	Symptoms, medications & exercise	6 (2.4 %)	2 (2%)	
	Symptoms, medications & others	4 (1.6 %)	0 (0%)	
	Symptoms, medications & study and research related	0 (0%)	2 (2%)	
	Symptoms, treatment options & nutrition	3 (1.2 %)	4 (4.1 %)	
	Symptoms, treatment options, nutrition & exercise	1 (0.4 %)	0 (0%)	
	Symptoms, treatment options & exercise	6 (2.4 %)	0 (0%)	
	Symptoms, treatment options, exercise & others	3 (1.2 %)	0 (0%)	
	Symptoms, nutrition & exercise	1 (0.4 %)	0 (0%)	
	Symptoms, nutrition & others	3 (1.2 %)	0 (0%)	
	Symptoms, medications, nutrition, exercise & others	1 (0.4 %)	0 (0%)	
	Symptoms, medications & treatment options	28 (11 %)	7 (7.1 %)	
	Symptoms, medications, treatment options, nutrition & exercise	35 (13. 8%)	6 (6.1 %)	
	Symptoms, medications, treatment options, nutrition, exercise & others	3 (1.2 %)	0 (0%)	
	Symptoms, medications,	3 (1.2 %)	0 (0%)	

	treatment options, nutrition, exercise & study and research related			
	Medications & treatment options	5 (2%)	6 (6.1%)	
	Medications & nutrition	3 (1.2%)	0 (0%)	
	Medications, treatment options & nutrition	0 (0%)	4 (4.1%)	
	Medications, treatment options, nutrition & exercise	2 (0.8%)	0 (0%)	
	Medications, nutrition & exercise	2 (0.8%)	0 (0%)	
	Nutrition & exercise	5 (2%)	2 (2%)	
	Treatment options, nutrition & exercise	3 (1.2%)	0 (0%)	
	Treatment options & exercise	4 (1.6%)	0 (0%)	
The way of the first learn about ChatGPT	**No**	0 (0%)	6 (6.1%)	**<0.001***
	Children	6 (2.4%)	3 (3.1%)	
	Family	4 (1.6%)	3 (3.1%)	
	Friends	96 (37.8%)	28 (28.6%)	
	News	11 (4.3%)	9 (9.2%)	
	Social media	111 (43.7%)	45 (45.9%)	
	Social media and friends	1 (0.4%)	0 (0%)	
	Trends	3 (1.2%)	0 (0%)	
	University	19 (7.5%)	1 (1%)	
	All of the above	3 (1.2%)	3 (3.1%)	
Believing that ChatGPT provides accurate medical information		3.15 ± 1.07	2.67 ± 1.48	**0.001***
		3 (2-4)	2 (1-4)	
ChatGPT's explanations are easy to understand		4.09 ± 1.02	3.28 ± 1.55	**<0.001***
		4 (3-5)	4 (1-5)	
ChatGPT gives balanced and unbiased health advice		3.62 ± 1.12	2.84 ± 1.45	**<0.001***
		3 (3-5)	4 (1-5)	
ChatGPT is more convenient than visiting a doctor for minor issues		2.92 ± 1.43	2.74 ± 1.7	0.344
		3 (1-4)	1.5 (1-4)	
Trusting ChatGPT's medical advice as much as information from Google		3.35 ± 1.3	2.76 ± 1.65	**<0.001***
		3 (2-5)	2.5 (1-5)	
ChatGPT's language makes me		3.45 ± 1.12	2.88 ± 1.6	**<0.001***

feel confident in its answers		3 (3-4)	2.5 (1-5)	
Believing that ChatGPT always uses up-to-date medical information		3.42 ± 1.08	2.95 ± 1.46	**0.002***
		3 (3-4)	3 (1-5)	
Feeling ChatGPT sometimes gives conflicting or confusing answers		3.19 ± 1.21	3.7 ± 1.15	**<0.001***
		3 (2-4)	4.5 (2-5)	
Ever acted on ChatGPT's medical advice without consulting a doctor		125 (49.2%)	18 (18.4%)	**<0.001***
Relying on ChatGPT in an emergency	**Yes**	51 (20.1%)	11 (11.2%)	**0.030***
	No	109 (42.9%)	37 (37.8%)	
	Not sure	94 (37%)	50 (51%)	
Always verifying ChatGPT's answers with official health websites or professionals		3.55 ± 1.25	3.84 ± 1.24	0.051
		4 (3-5)	4 (3-5)	
Aware that ChatGPT can sometimes make factual errors		4.08 ± 1.13	4.05 ± 1.26	0.808
		5 (3-5)	4 (3-5)	
Thinking ChatGPT should include clear warnings when giving medical information		4.18 ± 1.14	4.42 ± 1.09	0.083
		5 (3-5)	5 (4-5)	
Satisfaction with ChatGPT as a medical-information source	**Very unsatisfied**	8 (3.1%)	17 (17.3%)	**<0.001***
	Unsatisfied	41 (16.1%)	32 (32.7%)	
	Neutral	89 (35%)	30 (30.6%)	
	Satisfied	92 (36.2%)	15 (15.3%)	
	Very Satisfied	24 (9.4%)	4 (4.1%)	
Recommending ChatGPT to friends or family for health advice	**Yes**	178 (70.1%)	42 (42.9%)	**<0.001***
	No	41 (16.1%)	37 (37.8%)	
	Not sure	35 (13.8%)	19 (19.4%)	

Influence of Participant Characteristics

Gender-based comparisons are presented in **Table 9**. Females were significantly older than males ($P < 0.001$). Males were more likely to have a medical background and prior awareness of ChatGPT ($P = 0.013$ and 0.009, respectively). Significant differences were also observed in usage frequency, types of topics searched, and sources of awareness ($P < 0.001$ for all). Males reported higher perceived accuracy ($P = 0.004$), greater trust in ChatGPT compared to Google ($P = 0.008$), and were more likely to act on ChatGPT's advice without consulting a doctor ($P = 0.002$). Satisfaction levels were also significantly higher among males ($P = 0.004$). In contrast, females were significantly more likely to verify ChatGPT responses with official health sources ($P < 0.001$). No significant differences were observed between genders in several other perception domains or recommendation behaviour.

Table 9: Relation between gender and demographic data and AI knowledge level

		Gender		**P value**
		Male (n=88)	**Female (n=264)**	
Age (years)		31.26 ± 12.41	40.98 ± 12.31	**<0.001***
Education level	**Intermediate education**	0 (0%)	6 (2.3%)	0.244
	High school	20 (22.7%)	58 (22%)	
	Bachelor	51 (58%)	149 (56.4%)	
	Postgraduate	9 (10.2%)	34 (12.9%)	

	Diploma	3 (3.4%)	10 (3.8%)	
	Master	1 (1.1%)	0 (0%)	
	PhD	4 (4.5%)	4 (1.5%)	
	Assistant professor	0 (0%)	3 (1.1%)	
Medical background		75 (85.2%)	190 (72%)	**0.013***
Ever heard of ChatGPT or AI chatbots		86 (97.7%)	234 (88.6%)	**0.009***
How often they use ChatGPT for health-related questions	**Daily**	30 (34.1%)	25 (9.5%)	**<0.001***
	Occasionally	24 (27.3%)	112 (42.4%)	
	Rarely	24 (27.3%)	83 (31.4%)	
	Weekly	10 (11.4%)	44 (16.7%)	
Kind of topics they usually ask about	**Symptoms**	9 (10.2%)	24 (9.1%)	**<0.001***
	Medication	1 (1.1%)	18 (6.8%)	
	Treatment Options	1 (1.1%)	10 (3.8%)	
	Nutrition	1 (1.1%)	3 (1.1%)	
	Exercise	8 (9.1%)	3 (1.1%)	
	Study and research related	0 (0%)	11 (4.2%)	
	Others	3 (3.4%)	17 (6.4%)	
	Not used	0 (0%)	3 (1.1%)	
	Symptoms & medications	1 (1.1%)	24 (9.1%)	
	Symptoms & treatment options	11 (12.5%)	13 (4.9%)	
	Symptoms & nutrition	0 (0%)	3 (1.1%)	
	Symptoms & exercise	1 (1.1%)	2 (0.8%)	
	Symptoms & others	0 (0%)	3 (1.1%)	
	Symptoms, medications, treatment options & nutrition	5 (5.7%)	5 (1.9%)	
	Symptoms, medications, treatment options & exercise	1 (1.1%)	0 (0%)	
	Symptoms, medications & nutrition	0 (0%)	8 (3%)	
	Symptoms, medications, nutrition & exercise	0 (0%)	7 (2.7%)	
	Symptoms, medications, nutrition, exercise & others	0 (0%)	2 (0.8%)	

	Symptoms, medications & exercise	3 (3.4 %)	5 (1.9 %)	
	Symptoms, medications & others	0 (0%)	4 (1.5 %)	
	Symptoms, medications & study and research related	2 (2.3 %)	0 (0%)	
	Symptoms, treatment options & nutrition	0 (0%)	7 (2.7 %)	
	Symptoms, treatment options, nutrition & exercise	0 (0%)	1 (0.4 %)	
	Symptoms, treatment options & exercise	0 (0%)	6 (2.3 %)	
	Symptoms, treatment options, exercise & others	0 (0%)	3 (1.1 %)	
	Symptoms, nutrition & exercise	0 (0%)	1 (0.4 %)	
	Symptoms, nutrition & others	0 (0%)	3 (1.1 %)	
	Symptoms, medications, nutrition, exercise & others	0 (0%)	1 (0.4 %)	
	Symptoms, medications & treatment options	12 (13. 6%)	23 (8.7 %)	
	Symptoms, medications, treatment options, nutrition & exercise	18 (20. 5%)	23 (8.7 %)	
	Symptoms, medications, treatment options, nutrition, exercise & others	0 (0%)	3 (1.1 %)	
	Symptoms, medications, treatment options, nutrition, exercise & study and research related	3 (3.4 %)	0 (0%)	
	Medications & treatment options	1 (1.1 %)	10 (3.8 %)	

	Medications & nutrition	0 (0%)	3 (1.1%)	
	Medications, treatment options & nutrition	0 (0%)	4 (1.5%)	
	Medications, treatment options, nutrition & exercise	0 (0%)	2 (0.8%)	
	Medications, nutrition & exercise	0 (0%)	2 (0.8%)	
	Nutrition & exercise	6 (6.8%)	1 (0.4%)	
	Treatment options, nutrition & exercise	0 (0%)	3 (1.1%)	
	Treatment options & exercise	1 (1.1%)	3 (1.1%)	
The way of the first learn about ChatGPT	**No**	0 (0%)	6 (2.3%)	**<0.001***
	Children	0 (0%)	9 (3.4%)	
	Family	1 (1.1%)	6 (2.3%)	
	Friends	25 (28.4%)	99 (37.5%)	
	News	15 (17%)	5 (1.9%)	
	Social media	41 (46.6%)	115 (43.6%)	
	Social media and friends	0 (0%)	1 (0.4%)	
	Trends	0 (0%)	3 (1.1%)	
	University	6 (6.8%)	14 (5.3%)	
	All of the above	0 (0%)	6 (2.3%)	
Believing that ChatGPT provides accurate medical information		3.31 ± 0.99	2.93 ± 1.26	**0.004***
		3 (3-4)	3 (2-4)	
ChatGPT's explanations are easy to understand		4.08 ± 0.99	3.81 ± 1.3	0.081
		4 (3-5)	4 (3-5)	
ChatGPT gives balanced and unbiased health advice		3.62 ± 1.19	3.35 ± 1.28	0.083
		4 (3-5)	3 (2-5)	
ChatGPT is more convenient than visiting a doctor for minor issues		3.03 ± 1.55	2.82 ± 1.48	0.268
		3 (2-5)	3 (1-4)	
Trusting ChatGPT's medical advice as much as information from Google		3.55 ± 1.4	3.08 ± 1.4	**0.008***
		4 (2-5)	3 (2-4)	
ChatGPT's language makes me feel confident in its answers		3.5 ± 1.2	3.24 ± 1.3	0.095
		4 (3-4)	3 (2-4)	
Believing that ChatGPT always uses up-to-date medical information		3.43 ± 1.15	3.26 ± 1.21	0.257
		3 (3-4)	3 (3-4)	
Feeling ChatGPT sometimes gives conflicting or confusing answers		3.5 ± 1.04	3.25 ± 1.26	0.101
		3.5 (3-4)	3 (2-4)	
Ever acted on ChatGPT's medical advice without consulting a doctor		48 (54.5%)	95 (36%)	**0.002***
Relying on ChatGP	**Yes**	30 (34.1%)	32 (12.1%)	**<0.001***

T in an emergency	**No**	29 (33%)	117 (44.3%)	
	Not sure	29 (33%)	115 (43.6%)	
Always verifying ChatGPT's answers with official health websites or professionals		3.16 ± 1.36	3.78 ± 1.17	**<0.001***
		3 (2-4)	4 (3-5)	
Aware that ChatGPT can sometimes make factual errors		4.13 ± 1.25	4.06 ± 1.13	0.646
		5 (3-5)	4 (3-5)	
Thinking ChatGPT should include clear warnings when giving medical information		4.33 ± 1.06	4.21 ± 1.16	0.389
		5 (4-5)	5 (3-5)	
Satisfaction with ChatGPT as a medical-information source	**Very unsatisfied**	6 (6.8%)	19 (7.2%)	**0.004***
	Unsatisfied	16 (18.2%)	57 (21.6%)	
	Neutral	18 (20.5%)	101 (38.3%)	
	Satisfied	40 (45.5%)	67 (25.4%)	
	Very Satisfied	8 (9.1%)	20 (7.6%)	
Recommending ChatGPT to friends or family for health advice	**Yes**	62 (70.5%)	158 (59.8%)	0.068
	No	19 (21.6%)	59 (22.3%)	
	Not sure	7 (8%)	47 (17.8%)	

As shown in Table 10, participants with a medical background were significantly younger than those without (P < 0.001). They were more likely to have heard of ChatGPT (P = 0.009) and used it more frequently for health-related queries (P < 0.001). No significant association was found with educational level (P = 0.720).

Medical background was significantly associated with higher perceived accuracy (P = 0.014), better perception of balanced advice (P = 0.004), greater awareness of errors (P = 0.050), higher likelihood of acting on advice (P = 0.002), and higher satisfaction (P = 0.008). No significant differences were observed in understanding, convenience, trust compared to Google, confidence, up-to-date perception, conflicting responses, emergency use, verification behaviour, or recommendation.

Overall, participants with medical background showed greater trust, satisfaction, and behavioural reliance on ChatGPT.

Table 10: Relation between medical background and demographic data and AI knowledge level

		Medical background		**P value**
		No (n=87)	**Yes (n=265)**	
Age (years)		42.93 ± 11.42	37.12 ± 13.21	**<0.001***
Education level	**Intermediate education**	2 (2.3%)	4 (1.5%)	0.720
	High school	17 (19.5%)	61 (23%)	
	Bachelor	55 (63.2%)	145 (54.7%)	
	Postgraduate	8 (9.2%)	35 (13.2%)	
	Diploma	4 (4.6%)	9 (3.4%)	
	Master	0 (0%)	1 (0.4%)	
	PhD	1 (1.1%)	7 (2.6%)	
	Assistant professor	0 (0%)	3 (1.1%)	
Ever heard of ChatGPT or AI chatbots		73 (83.9%)	247 (93.2%)	**0.009***
How often they use	**Daily**	12 (13.8%)	43 (16.2%)	**<0.001***

ChatGPT for health-related questions	**Occasionally**	30 (34.5%)	106 (40%)	
	Rarely	41 (47.1%)	66 (24.9%)	
	Weekly	4 (4.6%)	50 (18.9%)	
The way of the first learn about ChatGPT	**No**	4 (4.6%)	2 (0.8%)	**0.016***
	Children	4 (4.6%)	5 (1.9%)	
	Family	4 (4.6%)	3 (1.1%)	
	Friends	33 (37.9%)	91 (34.3%)	
	News	5 (5.7%)	15 (5.7%)	
	Social media	35 (40.2%)	121 (45.7%)	
	Social media and friends	0 (0%)	1 (0.4%)	
	Trends	0 (0%)	3 (1.1%)	
	University	0 (0%)	20 (7.5%)	
	All of the above	2 (2.3%)	4 (1.5%)	
Believing that ChatGPT provides accurate medical information		2.75 ± 1.37	3.12 ± 1.13	**0.014***
		3 (2-4)	3 (3-4)	
ChatGPT's explanations are easy to understand		3.73 ± 1.35	3.93 ± 1.19	0.237
		4 (3-5)	4 (3-5)	
ChatGPT gives balanced and unbiased health advice		3.07 ± 1.39	3.53 ± 1.2	**0.004***
		3 (2-4)	4 (3-5)	
ChatGPT is more convenient than visiting a doctor for minor issues		2.84 ± 1.58	2.89 ± 1.48	0.815
		3 (1-4)	3 (1-4)	
Trusting ChatGPT's medical advice as much as information from Google		3.23 ± 1.52	3.19 ± 1.39	0.812
		3 (2-5)	3 (2-4)	
ChatGPT's language makes me feel confident in its answers		3.34 ± 1.33	3.3 ± 1.26	0.796
		3 (2.75-4.25)	3 (3-4)	
Believing that ChatGPT always uses up-to-date medical information		3.35 ± 1.33	3.29 ± 1.16	0.731
		3 (3-5)	3 (3-4)	
Feeling ChatGPT sometimes gives conflicting or confusing answers		3.18 ± 1.22	3.36 ± 1.21	0.243
		3 (2-4)	3 (3-4)	
Ever acted on ChatGPT's medical advice without consulting a doctor		23 (26.4%)	120 (45.3%)	**0.002***
Relying on ChatGPT in an emergency	**Yes**	18 (20.7%)	44 (16.6%)	0.202
	No	29 (33.3%)	117 (44.2%)	
	Not sure	40 (46%)	104 (39.2%)	
Always verifying ChatGPT's answers with official health websites or professionals		3.51 ± 1.28	3.66 ± 1.24	0.363
		4 (3-5)	4 (3-5)	
Aware that ChatGPT can sometimes make factual errors		3.84 ± 1.24	4.15 ± 1.13	**0.050***
		4 (3-5)	5 (3-5)	
Thinking ChatGPT should include clear warnings when giving medical information		4.16 ± 1.23	4.27 ± 1.1	0.449
		5 (3-5)	5 (4-5)	
Satisfaction with ChatGPT as a medical-	**Very unsatisfied**	11 (12.6%)	14 (5.3%)	**0.008***
	Unsatisfied	23 (26.4%)	50 (18.9%)	

information source	**Neutral**	31 (35.6%)	88 (33.2%)	
	Satisfied	20 (23%)	87 (32.8%)	
	Very Satisfied	2 (2.3%)	26 (9.8%)	
Recommending ChatGPT to friends or family for health advice	**Yes**	45 (51.7%)	175 (66%)	0.057
	No	25 (28.7%)	53 (20%)	
	Not sure	17 (19.5%)	37 (14%)	

Predictors of Trust, Satisfaction, and Recommendation

Multiple regression analysis for predictors of general public trust is shown in **Table 11**. The external independent reviewer total score (P < 0.001) and comparator site completeness (P = 0.004) were significant predictors.

Table 11: Predictors of general public trust

	Coefficient	**SE**	**t**	**P**	**r partial**	**r semi partial**
Quality Score of expert reviewer 1	0.000	0.059	0.007	0.994	0.001	0.001
Quality score of expert reviewer 2	0.018	0.045	0.395	0.694	0.049	0.035
External independent reviewer	0.226	0.051	4.411	**<0.001***	0.480	0.395
total score						
Comparator sites accuracy	0.119	0.218	0.544	0.589	0.067	0.049
Comparator sites completeness	-0.317	0.107	-2.948	**0.004***	-0.344	0.264
General public easy	-0.158	0.081	-1.951	0.055	-0.235	0.175
General public helpful	0.191	0.183	1.044	0.300	0.128	0.094

As shown in **Table 12**, significant predictors of satisfaction with ChatGPT included age (P < 0.001), medical background (P = 0.001), frequency of use (P < 0.001), perceived accuracy (P = 0.017), trust compared to Google (P = 0.021), and perception of conflicting information (P = 0.003).

Table 12: Predictors of satisfaction with ChatGPT

Independent variables	**Coefficient**	**SE**	**t**	**P**	**r partial**	**r semi partial**
Age (years)	-0.0199	0.0044	-4.3440	**<0.001***	-0.2235	0.201
Gender	0.122	0.1255	0.9799	0.328	0.0544	0.045
Education level	0.047	0.0477	0.9944	0.321	0.0555	0.046
Medical background	0.419	0.1199	3.5288	**0.001***	0.1933	0.164
Frequency of using ChatGP	-0.382	0.055	-6.8	**<0.001***	-0.3	0.320

T for health-related questions			97		58	
Believing that ChatGPT provides accurate medical information	0.115	0.048	2.404	**0.017***	0.133	0.112
Trusting ChatGPT's medical advice as much as information from Google	0.094	0.041	2.317	**0.021***	0.128	0.108
Always verifying ChatGPT's answers with official health websites or professionals	0.041	0.046	0.888	0.375	0.049	0.041
Feeling ChatGPT sometimes gives conflicting or confusing answers	-0.132	0.044	-3.001	**0.003***	-0.165	0.139
Aware that ChatGPT can sometimes make factual errors	-0.052	0.051	-1.022	0.308	-0.057	0.047

Table 12 demonstrates that medical background (P = 0.019), frequency of use (P < 0.001), and perceived accuracy (P = 0.015) were significant predictors of recommending ChatGPT to others.

Independent variables	**Coefficient**	**SE**	**t**	**P**	**r partial**	**r semi partial**
Age (years)	0.005	0.003	1.821	0.070	0.101	0.081
Gender	-0.051	0.086	-0.592	0.554	-0.033	0.026
Education level	-0.032	0.033	-0.985	0.325	-0.055	0.044
Medical background	-0.192	0.082	-2.351	**0.019***	-0.130	0.105
Frequency of using ChatGPT for health-related questions	0.445	0.038	11.650	**<0.001***	0.544	0.518
Believing that ChatGPT provides accurate medical information	-0.081	0.033	-2.448	**0.015***	-0.135	0.109
Trusting ChatGPT's medical advice as much as information from Google	0.033	0.028	1.182	0.238	0.066	0.053
Always verifying ChatGPT's answers with official health websites or	0.013	0.032	0.403	0.687	0.022	0.018

professionals						
Feeling ChatGPT sometimes gives conflicting or confusing answers	0.039	0.030	1.293	0.197	0.072	0.058
Aware that ChatGPT can sometimes make factual errors	0.014	0.035	0.403	0.687	0.022	0.018

Discussion

This investigation provides a distinctive contribution by combining systematic validation of AI-generated content with analysis of user behaviours, offering a thorough assessment of ChatGPT's role as a cardiovascular health information resource. The study systematically examined both public attitudes and expert appraisals of ChatGPT's effectiveness in delivering cardiovascular health education among 352 respondents. The results yield important insights into utilization patterns, accuracy perceptions, readability assessments, potential bias, trust dynamics, safety-related behaviours, and professional evaluations that demonstrate both concordance with and divergence from published literature.

Regarding the prevalence of ChatGPT use for medical information, 72.2% of participants reported using ChatGPT for health-related queries. This rate substantially exceeds that reported by Hosseini et al. (11), who documented that only 40% of respondents had used ChatGPT for healthcare-related queries, representing a difference of 32.2 percentage points. This discrepancy likely reflects the rapid adoption trajectory of generative AI tools in healthcare contexts over the intervening period. The accelerated uptake observed aligns with broader trends documented by Patel and Lam (1), who noted exponential growth in ChatGPT applications across various medical domains following its public release. Furthermore, Ayers et al. (5) observed increasing public comfort with seeking health information through digital platforms, a phenomenon intensified since the COVID-19 pandemic heightened reliance on virtual health resources. The 72.2% usage rate suggests that healthcare systems and regulatory bodies must urgently address the implications of widespread AI chatbot use for cardiovascular health information.

Concerning perceived accuracy of ChatGPT responses, participants rated accuracy at a moderate level with a mean score of 3.03 ± 1.2 out of 5 (60.6%). This finding demonstrates alignment with objective accuracy assessments in the literature. Andrew (12) conducted a meta-analysis of ChatGPT performance on cardiology board-style questions and reported a pooled accuracy of 58.64% (95% CI: 52.01%–65.13%), closely matching our participants' perception. Similarly, Skalidis et al. (13) documented 58.8% accuracy on European Exam in Core Cardiology questions, further supporting alignment with perceived accuracy ratings. However, our findings oppose higher accuracy rates reported in specialized contexts. Kassab et al. (14) demonstrated that GPT-4 achieved 90.3% accuracy on cardiovascular clinical vignettes from USMLE examinations, with 89.2% of explanations rated as accurate by experienced cardiologists. Gurbuz and Varis (15) reported 90% accuracy when evaluating ChatGPT against acute coronary syndrome guidelines, and Harskamp and De Clercq (16) documented 92% correct responses in cardiovascular trivia assessment. The disparity between user perceptions (60.6%) and higher objective accuracy rates (87–92%) suggests that public awareness of ChatGPT's capabilities may lag behind its actual performance in structured medical knowledge tasks.

It is essential to acknowledge the considerable variability observed among evaluators, most notably the substantially reduced completeness ratings assigned by the external reviewer. This divergence indicates that although ChatGPT responses may demonstrate surface-level accuracy, they potentially lack the thoroughness and detail required for rigorous clinical assessment.

The gap between internal expert evaluations and external reviewer judgments warrants consideration of potential assessment bias and differing evaluation standards. Internal reviewers consistently provided elevated scores across most quality dimensions, whereas the external reviewer systematically assigned lower ratings, with completeness showing the most pronounced difference. This pattern reveals that AI-generated content, despite demonstrating

general accuracy, may exclude essential information required for thorough patient education. For critical conditions such as myocardial infarction, information gaps may foster misconceptions or postpone appropriate care-seeking, reinforcing the importance of exercising caution and complementing AI-derived information with professional medical guidance. In suspected myocardial infarction, ChatGPT may support general education but must not replace emergency medical evaluation or physician consultation.

With respect to ease of understanding and readability, participants rated ease of understanding favourably with a mean score of 3.88 ± 1.23 out of 5 (77.6%). This finding aligns with evidence demonstrating ChatGPT's capacity to produce comprehensible health information. King et al. (17) showed that GPT-4 successfully improved the readability of institutional heart failure patient education materials to the 6th–7th grade level, meeting American Medical Association recommendations. Bhupathi et al. (18) reported 83.75% understandability on the Patient Education Materials Assessment Tool for heart failure content, aligning with our 77.6% rating. However, favourable perceptions must be interpreted alongside contradictory evidence regarding objective readability metrics. Singavarapu et al. (8) found that ChatGPT 3.5 produced content at a Flesch-Kincaid Grade Level of 11.99, substantially exceeding the recommended 6th-grade reading level. Olszewski et al. (10) reported even higher complexity, with ChatGPT responses requiring a Flesch-Kincaid Grade Level of 15.9, equivalent to college-level reading ability. The tension between favourable user perceptions (77.6%) and concerning objective readability measures (grade levels 11.99–15.9) suggests that users may conflate comprehensibility with other qualities such as response length, formatting, or conversational tone.

Regarding bias and neutrality perceptions, participants rated ChatGPT's neutrality at a moderate level with a mean score of 3.42 ± 1.26 out of 5 (68.4%). This moderate rating aligns with documented evidence of bias in large language models. Zhang et al. (19) demonstrated that ChatGPT exhibits gender and racial biases in acute coronary syndrome management recommendations, providing differential advice based on patient demographic characteristics. Fang et al. (20) examined AI-generated content and confirmed systematic biases in large language model outputs, though they noted comparatively lower bias levels than certain other content generation systems. Sarraju et al. (4) found generally appropriate cardiovascular disease prevention recommendations from ChatGPT but emphasized the need for clinician oversight to identify and correct potential biases. The moderate neutrality rating (68.4%) suggests that users maintain healthy skepticism regarding AI-generated health information, representing an important safeguard against uncritical acceptance of potentially biased content.

Concerning trust in information, the study revealed moderate trust levels with a mean score of 3.2 ± 1.42 out of 5 (64%). This finding aligns with comparative platform assessments. Niko et al. (21) found that ChatGPT demonstrated higher accuracy than Bing in responding to home blood pressure monitoring questions. Kozaily et al. (6) demonstrated that ChatGPT-3.5 provided appropriate answers to 90% of heart failure questions, which opposes our moderate trust rating, as objective appropriateness (90%) exceeds perceived trust (64%) by 26 percentage points. Dimitriadis et al. (22) reported that ChatGPT-4 achieved 100% accuracy in key areas including lifestyle advice, medication mechanisms, and symptom management for heart failure patients, which strongly opposes our moderate trust rating. King et al. (23) documented 98.1% appropriateness for GPT-3.5 on 107 heart failure questions with a hallucination rate of only 1.9%. The discrepancy between high objective appropriateness (90–100%) and moderate user trust (64%) may reflect public awareness of highly publicized AI failures or general caution regarding medical information from non-traditional sources. The elevated proportion of medically trained respondents may contribute to this disparity, indicating that knowledgeable users preserve critical evaluation despite satisfactory content quality.

Regarding perception of up-to-date information, participants rated currency with a mean score of 3.31 ± 1.19 out of 5 (66.2%), reflecting moderate confidence in the timeliness of responses. This finding aligns with documented concerns regarding knowledge cutoff limitations. Lee et al. (2) described that GPT-4's training data has temporal boundaries, meaning it cannot access information beyond its knowledge cutoff date. The concern is particularly relevant for cardiovascular medicine, where guidelines undergo regular updates. Sevinç et al. (7) demonstrated that while ChatGPT achieved high accuracy (95.14% for paid version; 91.21% for free version) on rare cardiovascular disease

information, the lowest-scoring category was "Treatment" at 90.7%, potentially reflecting rapidly evolving therapeutic recommendations.

With respect to response consistency, participants rated consistency with a mean score of 3.31 ± 1.21 out of 5 (66.2%). This perception opposes higher objective consistency rates documented in the literature. Kozaily et al. (6) reported 93% consistency for ChatGPT-3.5 responses to heart failure questions, exceeding perceived consistency by 26.8 percentage points. Zeljkovic et al. (24) evaluated ChatGPT-4's repeatability for atrial fibrillation information and found that responses contained consistent core medical information with no impact on correctness. However, our moderate rating aligns with concerns documented by Krügel et al. (25), who found that ChatGPT provides inconsistent moral advice that influences users' judgment, demonstrating context-dependent reliability particularly in ethically complex scenarios.

Concerning safety behaviours, 40.6% of participants reported acting on ChatGPT advice without consulting a healthcare professional. This finding aligns with documented safety concerns. Saenger et al. (26) reported a case of delayed diagnosis of transient ischemic attack caused by a patient's reliance on ChatGPT, illustrating real-world harms from unsupervised AI use. Pham et al. (27) documented suboptimal adherence to American Heart Association guidelines in ChatGPT responses to cardiac emergency scenarios. Shahid et al. (9) demonstrated that patients with inadequate health literacy were three times more likely to revisit the emergency department (OR: 3.0; 95% CI: 1.3–6.9, $p = 0.01$), suggesting that reliance on AI-generated information without professional verification may contribute to adverse outcomes.

Regarding emergency reliability, only 17.6% of participants indicated they would rely on ChatGPT in emergency situations, demonstrating appropriate caution. This low reliance rate strongly aligns with objective performance limitations. Pham et al. (27) found only 42% accuracy for bradycardia scenarios and 69% for cardiac arrest situations using American Heart Association Advanced Cardiovascular Life Support guidelines. Geneş and Deveci (28) found limitations in AI performance for time-critical cardiovascular decision-making. The low emergency reliance rate demonstrates appropriate risk perception, though continued education regarding AI limitations in acute settings remains essential.

Concerning awareness of potential errors, participants demonstrated high awareness with a mean score of 4.07 ± 1.16 out of 5 (81.4%). This finding strongly aligns with documented error rates and professional consensus. Lee et al. (2) described error rates ranging from 5% to 50% depending on clinical context. King et al. (23) reported a hallucination rate of 1.9% in heart failure responses. Soddu et al. (29) found that while 63.33% of responses achieved "mostly correct" ratings, 33.33% were only "moderately correct" and no responses achieved "completely correct" status. Andrew (12) documented substantial heterogeneity in ChatGPT accuracy across cardiology studies ($I^2 = 84.41\%$), validating user awareness that performance varies considerably.

Regarding the need for regulation and warning labels, participants strongly endorsed this with a mean score of 4.24 ± 1.13 out of 5 (84.8%), representing the highest-rated item. This finding strongly aligns with professional consensus and regulatory developments. Inam et al. (30) reviewed guidelines from leading cardiology journals and found that mandatory AI disclosure policies have been implemented across major publications. Lee et al. (2) emphasized that appropriate governance frameworks are essential for AI chatbot deployment in medical contexts. The convergence between public attitudes (84.8%) and professional guidelines represents an encouraging foundation for policy development.

Concerning expert evaluation of accuracy, expert evaluators rated ChatGPT's cardiovascular information highly, with scores ranging from 4.48 to 4.62 out of 5 (89.6%–92.4%). These ratings strongly align with favourable assessments in the literature. Goodman et al. (31) reported median accuracy of 5.5 out of 6 (91.7%), closely matching our range. Sevinç et al. (7) demonstrated accuracy of 91–95% when generating patient education resources for rare cardiovascular diseases. Kassab et al. (14) reported 89.2% of explanations rated as accurate by experienced cardiologists. Salihu et al. (32) found 77–90% agreement between ChatGPT recommendations and Heart Team decisions for severe aortic stenosis cases.

With respect to expert evaluation of completeness, expert evaluators rated completeness at approximately 3.9 out of 5 (78%), with the external reviewer rating substantially lower at 2.59 out of 5 (51.8%). This finding demonstrates mixed alignment

with existing literature. Goodman et al. (31) reported median completeness of 100%, which opposes our lower rating. However, our findings align with Anaya et al. (33), who found 78% actionability when comparing ChatGPT-3 with materials from leading cardiology institutes, precisely matching our completeness rating. Soddu et al. (29) found that 50.0% of responses scored "mostly complete" and 46.67% scored "adequate," reflecting moderate-to-high completeness consistent with our findings.

This study has several strengths, including a mixed methods design that combines expert evaluation with public perception assessment, involvement of multiple expert reviewers with reliability testing, and focus on a clinically important condition enabling detailed content analysis. However, limitations should be noted, including convenience sampling and a high proportion of participants with medical backgrounds (75.3%), which may limit generalizability. The sample was predominantly female (75%), and the cross-sectional design prevents causal inference. In addition, the study assessed English-language content only, expert ratings showed inter-rater variability, and results reflect a specific version of ChatGPT at one point in time.

The main finding is a clear mismatch between AI performance and user behaviour. Although ChatGPT showed generally acceptable accuracy, a considerable proportion of participants reported acting on its medical advice (40.6%), including use in emergency situations (17.6%). This indicates potential risk in real-world application, particularly when responses lack completeness or actionable detail. The discrepancy between acceptable accuracy, moderate user trust, and lower completeness suggests that users may rely on seemingly credible but insufficiently comprehensive information. Therefore, accuracy alone is not enough, and completeness and behavioural impact must also be considered when evaluating AI tools in healthcare.

Conclusions

This comprehensive evaluation reveals that ChatGPT demonstrates acceptable accuracy (expert ratings 89.6%–92.4%) and safety (81.0%–95.6%) for conveying core medical information about heart attacks, consistent with findings across cardiovascular medicine literature showing 77–95% appropriateness for clinical questions. The content is presented in an accessible manner that lay users find easy to understand (86.6%). However, completeness limitations, particularly highlighted by the external reviewer rating of 51.8%, indicate that ChatGPT should be viewed as a complement to, rather than replacement for, comprehensive authoritative health resources. The substantial proportion of users acting on AI advice without professional verification (40.6%) highlights a need for public education about appropriate AI use in health contexts.

Healthcare providers and public health practitioners should acknowledge ChatGPT's role in contemporary health information seeking while guiding patients toward balanced use that leverages AI accessibility while maintaining appropriate professional oversight for significant health concerns like suspected myocardial infarction. Continued monitoring of AI performance and user behaviour will be essential as these technologies evolve and become increasingly integrated into health communication ecosystems. ChatGPT may serve as a supplementary educational tool; however, it should not replace urgent medical evaluation or physician consultation for suspected myocardial infarction.

References

1. Patel SB, Lam K. ChatGPT: the future of discharge summaries? *Lancet Digit Health*. 2023;5(3):e107–e108. https://doi.org/10.1016/S2589-7500(23)00021-3
2. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388(13):1233–1239. https://doi.org/10.1056/NEJMsr2214184
3. Rodriguez F, Ngo S, Baird G, Balla S, Miles R, Garg M. Readability of online patient educational materials for coronary artery calcium scans and implications for health disparities. *J Am Heart Assoc*. 2020;9(18):e017372. https://doi.org/10.1161/JAHA.120.017372
4. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA*. 2023;329(10):842–844. https://doi.org/10.1001/jama.2023.1044

5. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183(6):589–596. https://doi.org/10.1001/jamainternmed.2023.1838
6. Kozaily E, Geagea M, Akdogan ER, Atkins J, Elshazly MB, Guglin M, et al. Accuracy and consistency of online large language model-based artificial intelligence chat platforms in answering patients' questions about heart failure. *Int J Cardiol*. 2024;408:132115. https://doi.org/10.1016/j.ijcard.2024.132115
7. Sevinç S, Candemir M, Yamak BA, Kızıltunç E, Sezenöz B, Şahin OB, et al. An academic evaluation of ChatGPT's ability and accuracy in creating patient education resources for rare cardiovascular diseases. *Sci Rep*. 2025;15(1):25929. https://doi.org/10.1038/s41598-025-11567-w
8. Singavarapu J, Khemlani A, Jacobs M, Berglas E, Lazar J, Kabarriti A. Evaluating the quality of cardiovascular disease information from AI chatbots: a comparative study. *Cureus*. 2025;17(7). https://doi.org/10.7759/cureus.88085
9. Shahid R, Shoker M, Chu LM, Frehlick R, Ward H, Pahwa P. Impact of low health literacy on patients' health outcomes: a multicenter cohort study. *BMC Health Serv Res*. 2022;22(1):1148. https://doi.org/10.1186/s12913-022-08527-9
10. Olszewski R, Watros K, Mańczak M, Owoc J, Jeziorski K, Brzeziński J. Assessing the response quality and readability of chatbots in cardiovascular health, oncology, and psoriasis: a comparative study. *Int J Med Inform*. 2024;190:105562. https://doi.org/10.1016/j.ijmedinf.2024.105562
11. Hosseini M, Gao CA, Liebovitz DM, Carber AM, Tiu C, Schwartz J, et al. An exploratory survey about using ChatGPT in education, healthcare, and research. *PLoS One*. 2023;18(10):e0292216. https://doi.org/10.1371/journal.pone.0292216
12. Andrew A. Accuracy of ChatGPT in answering cardiology board-style questions. *J Educ Eval Health Prof*. 2025;22. https://doi.org/10.3352/jeehp.2025.22.9
13. Skalidis I, Cagnina A, Fournier S. Performance of artificial intelligence in answering cardiovascular textual questions. *Eur Heart J Digit Health*. 2023;4(5):364–365. https://doi.org/10.1093/ehjdh/ztad042
14. Kassab J, Massad C, Kapadia V, El Hajjar AH, El Dahdah J, Chedid El Helou M, et al. Performance evaluation of ChatGPT 4.0 on cardiovascular clinical cases from the USMLE Step 2CK and Step 3 of the National Board of Medical Examiners. *Circulation*. 2023;148(Suppl_1):A16722. https://doi.org/10.1161/circ.148.suppl_1.16722
15. Gurbuz DC, Varis E. Is ChatGPT knowledgeable of acute coronary syndromes and pertinent European Society of Cardiology guidelines? *Minerva Cardiol Angiol*. 2024;72(4):299–303. https://doi.org/10.23736/S2724-5683.24.06517-7
16. Harskamp RE, De Clercq L. Performance of ChatGPT as an AI-assisted decision support tool in medicine: a proof-of-concept study for interpreting symptoms and management of common cardiac conditions (AMSTELHEART-2). *Acta Cardiol*. 2024;79(4):358–366. https://doi.org/10.1080/00015385.2024.2303528
17. King RC, Samaan JS, Haquang J, Bharani V, Margolis S, Srinivasan N, et al. Improving the readability of institutional heart failure-related patient education materials using GPT-4: observational study. *JMIR Cardio*. 2025;9:e68817. https://doi.org/10.2196/68817
18. Bhupathi M, Kareem JM, Mediboina A, Janapareddy K. Assessing information provided by ChatGPT: heart failure versus patent ductus arteriosus. *Cureus*. 2025;17:e86365. https://doi.org/10.7759/cureus.86365
19. Zhang A, Yuksekgonul M, Guild J, Zou J. ChatGPT exhibits gender and racial biases in acute coronary syndrome management. *medRxiv*. 2023.

https://doi.org/10.1101/2023.11.14.23298525
20. Fang X, Che S, Mao M, Zhang H, Liu Y. Bias of AI-generated content: an examination of news produced by large language models. *Sci Rep*. 2024;14:5224. https://doi.org/10.1038/s41598-024-55686-2
21. Niko MM, Karbasi Z, Kazemi M, Zahmatkeshan M. Comparing ChatGPT and Bing in response to the home blood pressure monitoring knowledge checklist. *Hypertens Res*. 2024;47(6):1401–1409. https://doi.org/10.1038/s41440-024-01624-8
22. Dimitriadis F, Alkagiet S, Tsigkriki L, Kleitsioti P, Sidiropoulos G, Efstratiou D, et al. ChatGPT and patients with heart failure. *Angiology*. 2025;76(8):796–801. https://doi.org/10.1177/00033197241238403
23. King RC, Samaan JS, Yeo YH, Mody B, Lombardo DM, Ghashghaei R. Appropriateness of ChatGPT in answering heart failure related questions. *Heart Lung Circ*. 2024;33:1314–1318. https://doi.org/10.1016/j.hlc.2024.03.005
24. Zeljkovic I, Novak M, Jordan A, Lisicic A, Nemeth-Blažić T, Pavlovic N, et al. Evaluating ChatGPT-4's correctness in patient-focused informing and awareness for atrial fibrillation. *Heart Rhythm O2*. 2025;6(1):58–63. https://doi.org/10.1016/j.hroo.2024.10.005
25. Krügel S, Ostermaier A, Uhl M. ChatGPT's inconsistent moral advice influences users' judgment. *Sci Rep*. 2023;13:4569. https://doi.org/10.1038/s41598-023-31341-0
26. Saenger JA, Hunger J, Boss A, Richter J. Delayed diagnosis of a transient ischemic attack caused by ChatGPT. *Wien Klin Wochenschr*. 2024;136(7–8):236–238. https://doi.org/10.1007/s00508-024-02329-1
27. Pham C, Govender R, Tehami S, Wong J, Kwan B. ChatGPT's performance in cardiac arrest and bradycardia simulations using the American Heart Association's advanced cardiovascular life support guidelines: exploratory study. *J Med Internet Res*. 2024;26:e55037. https://doi.org/10.2196/55037
28. Geneş M, Deveci B. A clinical evaluation of cardiovascular emergencies: a comparison of responses from ChatGPT, emergency physicians, and cardiologists. *Diagnostics*. 2024;14(23):2731. https://doi.org/10.3390/diagnostics14232731
29. Soddu M, De Vito A, Madeddu G, Nicolosi B, Provenzano M, Ivziku D, et al. Assessing the accuracy, completeness and safety of ChatGPT-4o responses on pressure injuries in infants: clinical applications and future implications. *Nurs Rep*. 2025;15(4):130. https://doi.org/10.3390/nursrep15040130
30. Inam M, Sheikh S, Minhas AMK, Fatima K, Nasir K, Virani SS. A review of top cardiology and cardiovascular medicine journal guidelines regarding the use of generative artificial intelligence tools in scientific writing. *Curr Probl Cardiol*. 2024;49(2):102387. https://doi.org/10.1016/j.cpcardiol.2024.102387
31. Goodman RS, Patrinely JR, Stone CA, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and reliability of chatbot responses to physician questions. *JAMA Netw Open*. 2023;6(10):e2336483. https://doi.org/10.1001/jamanetworkopen.2023.36483
32. Salihu A, Meier D, Noirclerc N, Windecker S. A study of ChatGPT in facilitating Heart Team decisions on severe aortic stenosis. *EuroIntervention*. 2024;20(5):e496–e503. https://doi.org/10.4244/EIJ-D-23-00643
33. Anaya F, Prasad R, Bashour M, Yaghmour R, Alameh A, Balakumaran K. Evaluating ChatGPT platform in delivering heart failure educational material: a comparison with the leading national cardiology institutes. *Curr Probl Cardiol*. 2024;49:102797. https://doi.org/10.1016/j.cpcardiol.2024.102797

Acknowledgments

The authors acknowledge all participants who contributed their time and perspectives to this research. We also thank the expert reviewers who provided careful evaluation of the ChatGPT-generated responses.

Conflicts of Interest

The authors declare no conflicts of interest.